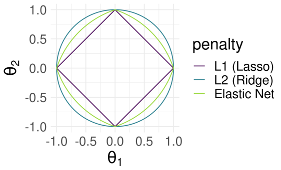
Introduction to Machine Learning

Elastic Net and Regularization for GLMs

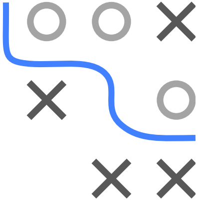


Learning goals

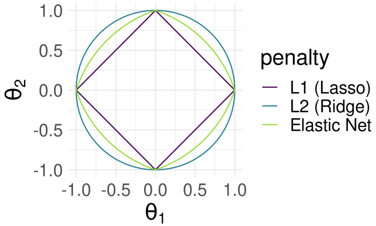
- Know the elastic net as compromise between ridge and lasso regression
- Know regularized logistic regression

ELASTIC NET

Elastic Net ▸ Zou and Hastie, 2005 combines the $L1$ and $L2$ penalties:

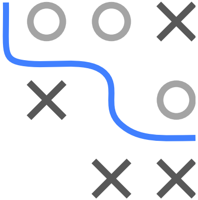


$$
\begin{aligned}
\mathcal{R}_{\text{elnet}}(\boldsymbol{\theta}) &= \sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \lambda_1 \|\boldsymbol{\theta}\|_1 + \lambda_2 \|\boldsymbol{\theta}\|_2^2 \\
&= \sum_{i=1}^{n} (y^{(i)} - \boldsymbol{\theta}^\top \mathbf{x}^{(i)})^2 + \lambda \left((1-\alpha) \|\boldsymbol{\theta}\|_1 + \alpha \|\boldsymbol{\theta}\|_2^2 \right), \; \alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}, \lambda = \lambda_1 + \lambda_2
\end{aligned}
$$



- Correlated features tend to be either selected or zeroed out together.
- Selection of more than n features possible for $p > n$.

ELASTIC NET



Simulating 50 data sets with 100 observations each for two coefficient settings:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \epsilon, \quad \epsilon \sim N(0, 1)$$

Ridge performs better for correlated features:

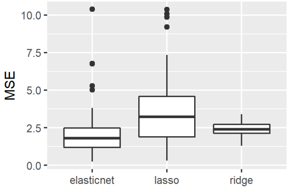
$$\boldsymbol{\theta} = (\underbrace{2, \ldots, 2}_{5}, \underbrace{0, \ldots, 0}_{5})$$

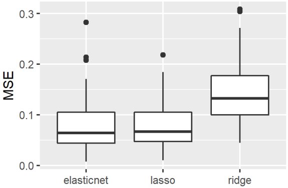
$\text{corr}(\mathbf{X}_i, \mathbf{X}_j) = 0.8^{|i-j|}$ for all i and j

Lasso performs better for sparse truth/no correlation:

$$\boldsymbol{\theta} = (2, 2, 2, \underbrace{0, \ldots, 0}_{7})$$

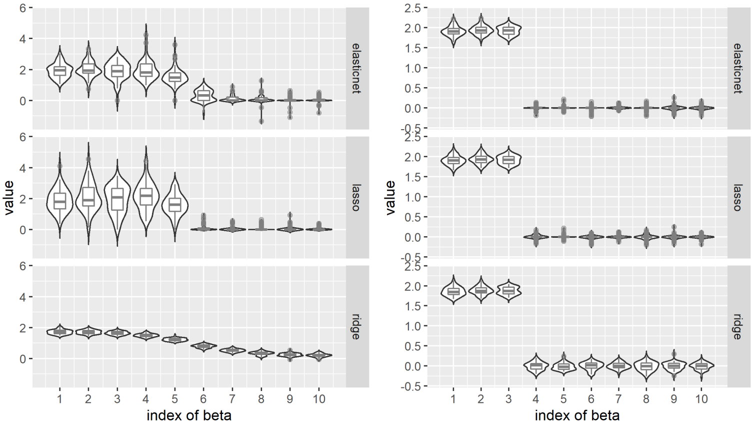
$\text{corr}(\mathbf{X}_i, \mathbf{X}_j) = 0$ for all $i \neq j$, otherwise 1





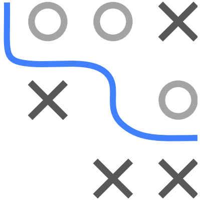
$\Longrightarrow$ Elastic Net handles both cases well

ELASTIC NET



LHS: ridge can not perform variable selection compared to lasso/E-Net. Lasso more often ignores relevant features than E-Net (longer tails in violin plot).
RHS: ridge estimates of noise features hover around 0 while lasso/E-Net produce 0s.

REGULARIZED LOGISTIC REGRESSION



Regularizers can be added very flexibly to basically any model which is based on ERM.

Hence, we can construct, e.g., $L1$- or $L2$-penalized logistic regression to enable coefficient shrinkage and variable selection in this model class.

$$
\begin{aligned}
\mathcal{R}_{\text{reg}}(\boldsymbol{\theta}) &= \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) + \lambda \cdot J(\boldsymbol{\theta}) \\
&= \sum_{i=1}^{n} \log\left[1 + \exp\left(-2y^{(i)} f\left(\mathbf{x}^{(i)} \;\middle|\; \boldsymbol{\theta}\right)\right)\right] + \lambda \cdot J(\boldsymbol{\theta})
\end{aligned}
$$

REGULARIZED LOGISTIC REGRESSION

We fit a logistic regression model using polynomial features for x_1 and x_2 with maximum degree of 7. We add an $L2$ penalty. We see for

- $\lambda = 0$: The unregularized model seems to overfit.
- $\lambda = 0.0001$: Regularization helps to learn the underlying mechanism.
- $\lambda = 1$: The real data-generating process is captured very well.